

# **Universidad de Buenos Aires**

## **Facultad de Ingeniería**



**Ciencia de Datos - Cátedra Martinelli**

**Trabajo Práctico N.º 1: “Análisis Exploratorio con Pandas”**

**2do Cuatrimestre 2025**

Alumno:

- Castro, Martín

# Introducción

En el presente trabajo se llevará a cabo un análisis exploratorio del conjunto de datos provisto por la cátedra. El propósito principal es desarrollar y aplicar habilidades orientadas a la exploración, explicación y visualización de la información de manera efectiva. A lo largo del informe se detallarán las decisiones adoptadas durante el proceso, así como las hipótesis planteadas y las conclusiones obtenidas a partir del análisis de los datos.

# Desarrollo

Como primer paso, se realizó la [estandarización](#) de los archivos CSV a utilizar. Este proceso consistió en transformar las columnas de tipo *string*, convirtiendo su contenido a mayúsculas o minúsculas, según resultara más conveniente, y eliminando los espacios en blanco adicionales que podrían generar inconvenientes en el desarrollo del trabajo.

En relación a los *outliers*, se decidió no descartarlos, ya que no se identificaron registros significativamente alejados de la media que pudieran afectar el análisis.

Asimismo, en ciertas consultas específicas fue necesario estandarizar columnas de otros tipos de datos, como aquellas de tipo *datetime*.

Finalmente, con el fin de comprender en mayor profundidad el dataset, se llevaron a cabo las consultas sugeridas por la cátedra:

## Consulta 1

¿Cuál es el estado que más descuentos tiene en total? y en promedio? Supongan que de una dirección del estilo: 3123 Alan Extension Port Andrea, MA 26926, “MA” es el estado. ([enlace](#))

Hipótesis:

- 1) Solo se consideraron las ‘orders’ con estado ‘completed’, ya que son las que completan todo el flujo del proceso: compra y recepción del pedido.
- 2) Se asume que todas las ‘shipping\_address’ contienen un código de estado seguido de un número (pueden tener otra información pero como mínimo debe tener esto), es decir, son de la forma:

*Estado número*

Esto es necesario para poder parsear las direcciones.

- 3) Se descartaron las filas con *shipping\_address* nulo ya que se consideró que no se sabía la dirección real o no se cargó el dato.
- 4) Se descartaron las filas con '*shipping\_address*' = "undefined" dado que los estados eran de USA y no existe un estado llamado "undefined". Se consideró que las filas con este valor hacían referencia a que no se tenía el dato o no se cargo.

Conclusiones:

Como puede verse en la imagen 1, el estado que tiene mayor cantidad de descuento en total es **TE** que supera ampliamente a **SC**, que ocupa el segundo lugar en esta categoría.

Esto se debe a que la cantidad de ventas en **TE** es similar a la diferencia que se llevan (casi cinco veces) en cantidad de ventas, por lo que podemos intuir que el estado **TE** no tiene grandes descuentos comparados a los demás sino que tiene una mayor cantidad de *orders*.

State	Descuento		Cantidad de ordenes
	Total	Promedio	Total
te	516806	12.437258	608104
sc	109029	12.400409	128520
it	82669	12.473563	97208
ap	27686	12.645974	32579
aa	27395	12.595908	32245

Imagen 1

Por otra parte, el estado con mayor descuento promedio es **HI**. Se puede observar que los estados con mayor promedio son aquellos que no tienen tantas *orders* (el top 5 tiene 30000 en promedio).

Otro punto a notar es que **TE** tenía un promedio de 12.437 mientras que **HI** tiene un promedio de 12.916 por lo que la diferencia no es significativa.

Esto se puede enfatizar en la imagen 3 donde vemos que los promedios rondan entre 12,4 y 12,6 con 2 outliers por fuera de estos valores.

State	Descuento		Cantidad de ordenes
	Total	Promedio	Total
hi	26412	12.916488	31070
dc	26494	12.761307	31222
vi	25973	12.738349	30582
ut	26009	12.725063	30582
tx	26376	12.689184	30940

Imagen 2

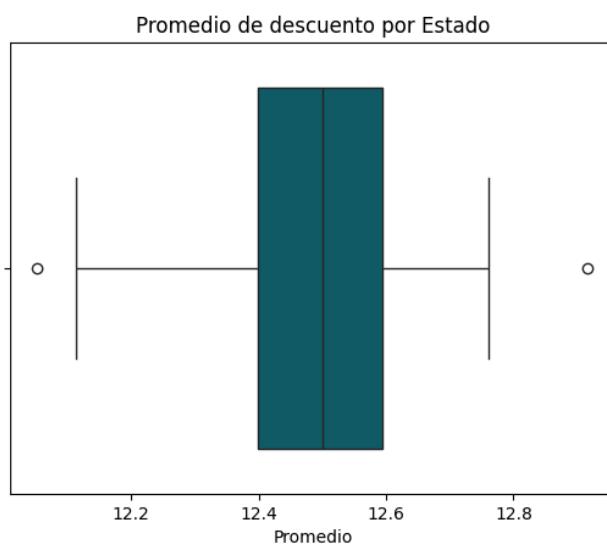


Imagen 3

Aclaración: Sí bien la consulta pedía el top 1, se analizaron más puestos del ranking ya que tenía información valiosa.

## Consulta 2

¿Cuáles son los 5 códigos postales más comunes para las órdenes con estado 'Refunded'? ¿Y cuál es el nombre más frecuente entre los clientes de esas direcciones? ([enlace](#))

Hipótesis:

- 1) Solo se consideraron las 'orders' con estado 'refunded' ya que lo pedía la consigna.
- 2) Eliminamos los códigos postales nulos ya que no nos aportaba información para la consulta que queremos hacer.

- 3) Eliminamos los nombres nulos por el mismo motivo. Si lo dejamos podría alterar los resultados.
- 4) Eliminamos los nombre *UNDEFINED* porque damos por sentado que no se llama nadie de esa forma y que es una forma de indicar que no se tiene esa información.

Conclusión:

Podemos ver en la imagen 4 que hay un cuádruple empate entre los códigos postales que más órdenes hicieron.

Un dato a notar es que el quinto puesto lo pueden ocupar distintos códigos ya que hay varios de estos con frecuencia 6. La elección de cual mostrar la hace la operación *sort\_values*.

top_5		
	postal_code	frecuencia
0	72397	7
1	15221	7
2	14655	7
3	14025	7
4	25226	6

Imagen 4

En la siguiente imagen podemos observar que hay un triple empate entre las personas que más paquetes pidieron a los códigos postales del top 5.

first_name	cantidad de compras	cantidad de codigos postales diferentes
DEBBIE	5	1
JESSICA	5	1
JOHN	5	1
CATHY	4	1
SHAWN	3	1
DENNIS	2	1
HARRY	2	1
LORI	2	1
TIFFANY	2	1
SANDRA	2	1
JULIE	1	1
VANESSA	1	1

Imagen 5

Las personas dentro del ranking pidieron a un único código postal, por lo que sumado a que la cantidad de órdenes no es muy grande, podemos pensar que se tratan de compras particulares y no de empresas o instituciones.

Aclaración: Nuevamente solamente se pedía el nombre con más apariciones dentro de los códigos postales del ranking pero aportaba un mayor valor para el análisis dejar la tabla completa.

## Consulta 3

Para cada tipo de pago y segmento de cliente, devolver la suma y el promedio expresado como porcentaje, de clientes activos y de consentimiento de marketing. Se valora que el output de la consulta tenga nombres claros y en español. ([enlace](#))

Hipótesis:

- 1) El promedio expresado como porcentaje se calcula como:  $mean() * 100$
- 2) Los clientes activos representan clientes que han realizado alguna compra en el último tiempo
- 3) El consentimiento de marketing representa que un usuario ha aceptado que se le envíe publicidad o se usen sus datos para cuestiones de marketing.
- 4) El consentimiento de marketing por método de pago, por ejemplo con tarjeta de crédito, representa que acepta que le envíe publicidad sobre promociones (o similar) con ese medio de pago
- 5) Los clientes activos por método de pago representan que en el último tiempo han pagado con ese método o tienen alguna tarjeta asociada (en el caso de pago con tarjeta)

Conclusión:

Se puede observar que hay un gran porcentaje de clientes activos en promedio (casi del 90%).

Por otro lado, el consentimiento de marketing tiene una buena aceptación (70%).

payment_method	customer_segment	Cliente activo		Consentimiento de marketing	
		Suma	Promedio porcentual	Suma	Promedio porcentual
bank transfer	budget	114885	89.927438	89153	69.785445
	premium	116360	89.822763	90546	69.895943
	regular	350102	89.934162	273106	70.155438
	undefined	19415	89.614586	15328	70.750058
cash on delivery	budget	114670	89.885086	88842	69.639582
	premium	117736	90.063186	91760	70.192617
	regular	350185	89.959411	273204	70.183677
	undefined	19209	89.606755	15052	70.215049
credit card	budget	113910	89.873368	88255	69.631938
	premium	116702	89.842643	91178	70.193078
	regular	350004	89.971158	273322	70.259474
	undefined	19490	89.592719	15155	69.665349
debit card	budget	113869	89.785056	88403	69.705261
	premium	117625	89.794875	91479	69.835029
	regular	351591	90.019049	273540	70.035384
	undefined	19451	89.623554	15060	69.391328
digital wallet	budget	114641	89.867990	89114	69.857172
	premium	117553	89.807097	91429	69.849116
	regular	351015	89.874565	274082	70.176490
	undefined	19328	89.809953	15038	69.875935
paypal	budget	114113	89.770761	88744	69.813399
	premium	117467	89.906242	91581	70.093758
	regular	351110	89.972146	274019	70.217531
	undefined	19479	89.748433	15250	70.263546
undefined	budget	22611	89.758247	17538	69.620102
	premium	23212	89.941104	18172	70.412275
	regular	69234	90.106200	53872	70.112968
	undefined	3705	89.341693	2854	68.820834

Imagen 6

Haciendo hincapié en los clientes activos podemos ver que la mayoría son *regulares* y no destaca ningún método de pago en particular.

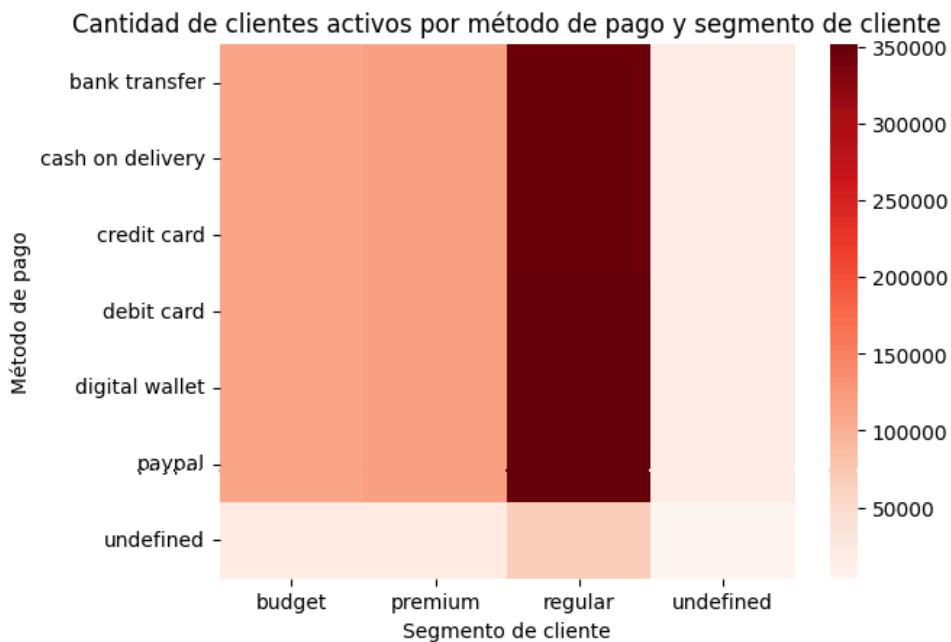


Imagen 7

El consentimiento de marketing también es mayormente aceptado por los clientes *regulares* aunque también hay muchos usuarios *budget* y *premium* que también dieron su consentimiento.

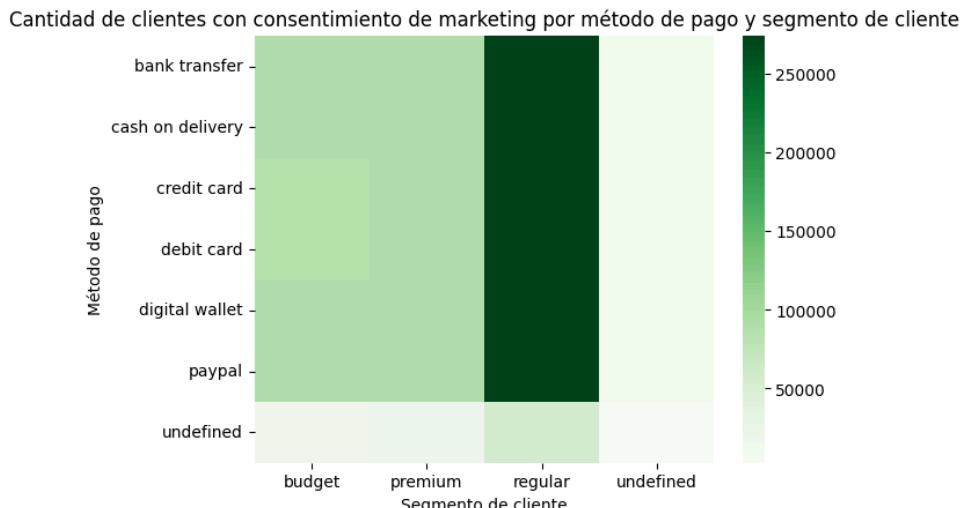


Imagen 8

## Consulta 4

Para los productos que contienen en su descripción la palabra “*stuff*” (sin importar mayúsculas o minúsculas), calcular el peso total de su inventario agrupado por marca, mostrar sólo la marca y el peso total de las 5 más pesadas. ([enlace](#))

Hipótesis:

- 1) Se quitan las filas las cuales tengan ‘*brand*’ nula ya que lo que buscamos en esta consulta es ver que marcas son las que mayor peso en stock tienen.
- 2) Se quitan las filas la cuales tengan ‘*brand*’ = “*undefined*” ya que consideramos que esta no es una marca, si no que faltan cargar esos datos o no se saben
- 3) Se quitan las filas que tengan la cantidad de stock o peso nulas ya que no nos aportan para calcular lo pedido.
- 4) Se conservaron las palabras que contienen “*stuff*” o alguna de sus variaciones como por ejemplo “*stuffs*”, ya que se interpretó que esto pedía la consigna.
- 5) Lo que pide la consulta es calcular el peso del inventario por marca. El cálculo se obtiene haciendo: peso de producto \* stock del producto (esto por cada fila donde esté esa marca)

## Conclusión:

Se puede observar que 3M es la marca que mayor peso tiene en cuanto stock en kg esto puede deberse al rubro que manejan mientras que el top 3 lo completan Adidas y Nike, dos empresas de ropa que si bien no guardan artículos de tanto peso podemos intuir que guardan una mayor cantidad de productos.

total_stock_weight	
brand	
3M	2250899.66
ADIDAS	1923907.88
NIKE	1783569.89
HASBRO	1714411.23
WAYFAIR	1666836.35

Imagen 9

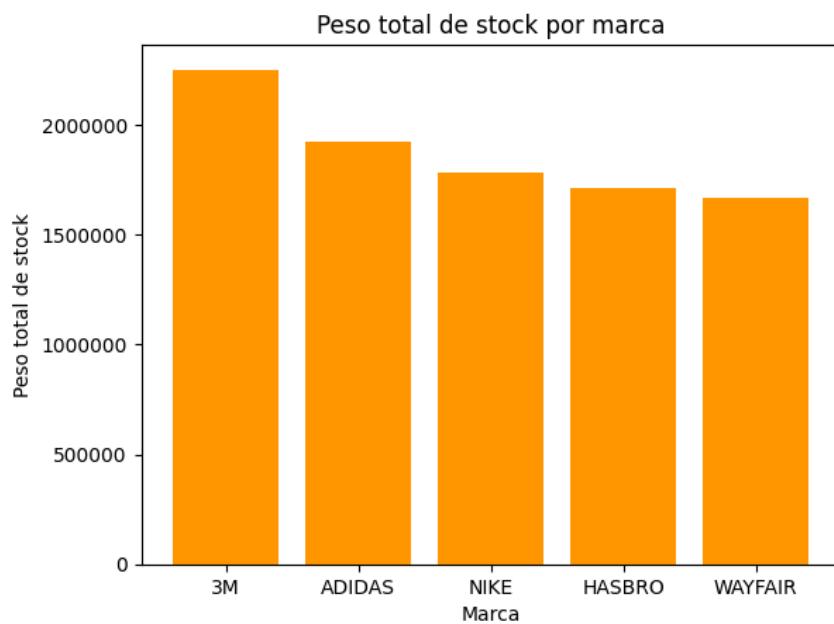


Imagen10

Al fijarnos en la imagen 10 vemos que esto no es así. 3M tiene mayor cantidad de artículos que los demás.

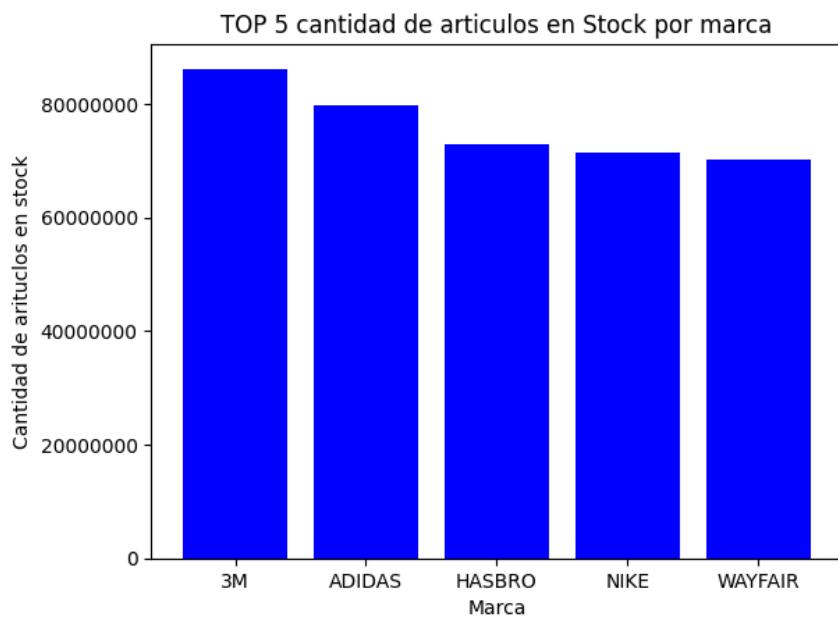


Imagen 11

La diferencia que podemos encontrar es entre Nike y Hasbro, por lo que aquí sí podemos decir que Nike guarda artículos más pesados, en promedio, que Hasbro.

Sí bien tomamos como hipótesis que no íbamos a considerar el '*brand*' “*undefined*” para la consulta, se hizo la prueba para ver cuántos artículos de marca sin definir había y fue el top 1 con diferencia. Por lo que podemos concluir que hay muchos artículos que no se sabe a qué marca pertenecen o que no se tiene ese dato.

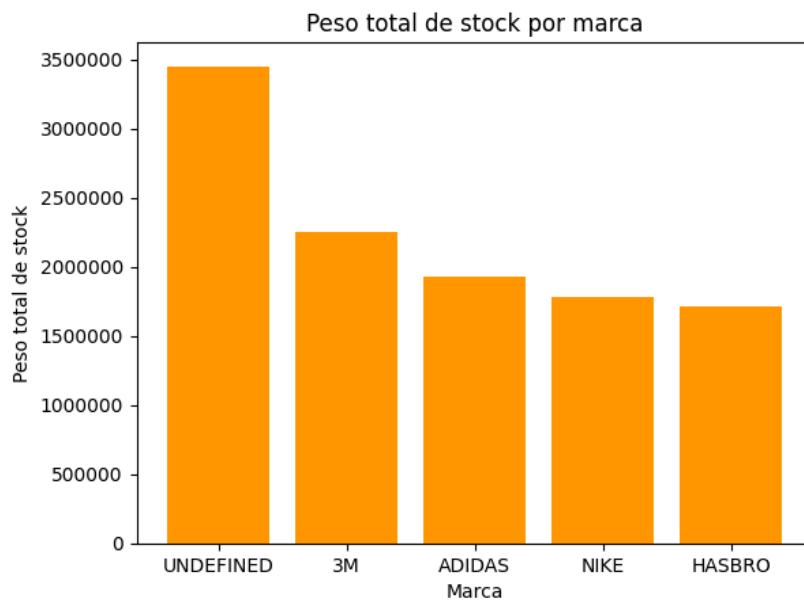


Imagen 12

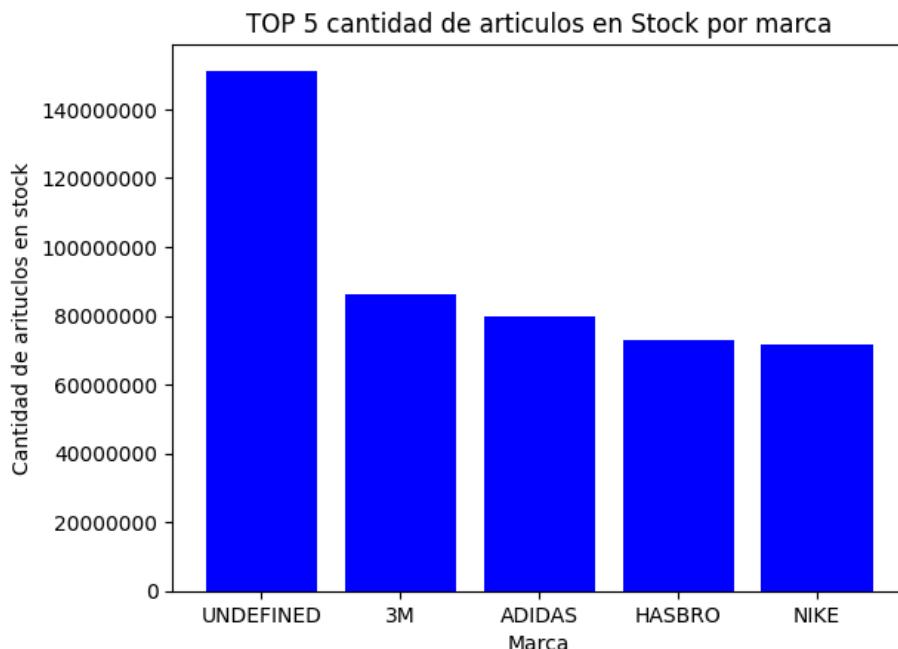


Imagen 13

## Consulta 5

La consulta consiste en ver la cantidad de ventas en el año 2024. Esto puede darnos una idea de si tenemos una tendencia de mayor o menor cantidad de ventas y en qué épocas del año. ([enlace](#))

Hipótesis:

- 1) Los valores con ‘order\_date’ nulos no serán considerados ya que no se sabe su fecha
- 2) Los valores con ‘order\_date’ que no tengan ningún formato de DateTime no serán considerados ya que no podemos parsearlos

Conclusión:

En esta consulta podemos ver que en la columna ‘order\_date’ había DateTime con distintos formatos lo que nos obliga al uso de `format = "mixed"` en la transformación de la columna a DateTime.

Yendo al análisis del resultado, encontramos un gráfico sorprendente ya que hay una tendencia de crecimiento durante todo el año 2024.

En el gráfico no se ven picos pronunciados por lo que tuvo un crecimiento estable de enero a diciembre.

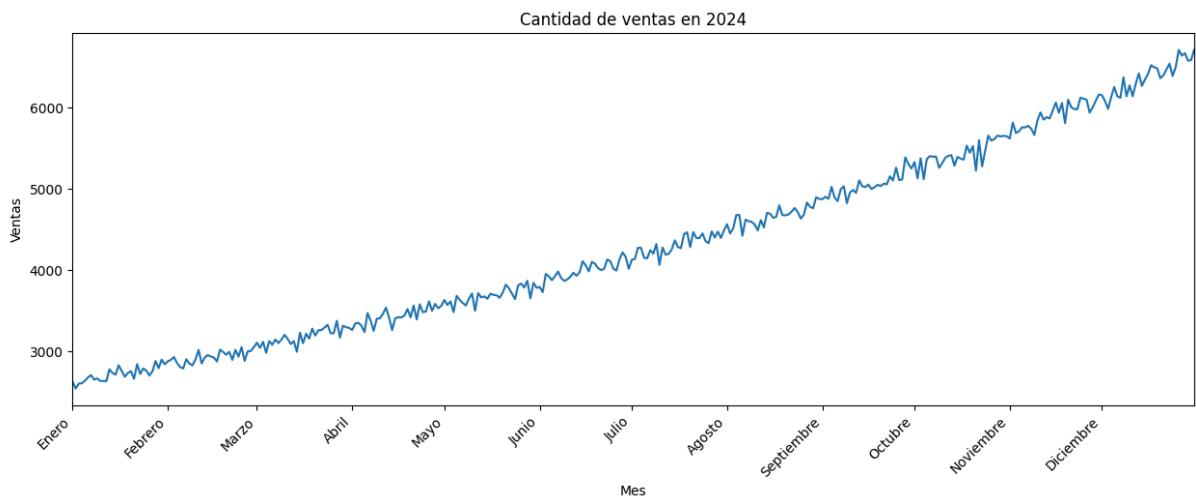


Imagen 14

## Consulta 6

En esta consulta se quiere comparar la cantidad de ventas en el Black Friday y Cyber Monday VS la cantidad de ventas una semana antes de navidad. La motivación de esta consulta es ver si las personas compran más en los días de mayores ofertas o dejan las compras para último momento. ([enlace](#))

Hipótesis:

- 1) Los valores nulos de la columna ‘order\_date’ no serán considerados ya que no se sabe en qué fecha fue la orden por lo que no puede agregarse.
- 2) Los valores de la columna ‘order\_date’ que no tengan ningún formato de DateTime serán eliminados.
- 3) Considero que las órdenes de pedidos son todas de USA ya que tienen estados de ese país.
- 4) El Black Friday fue el 29 de noviembre y el Cyber Monday el 2 de diciembre.

Conclusión:

Podemos observar que mucha gente no hizo compras en los días que debiera haber mayores ofertas y la gran mayoría hizo sus compras navideñas una semana antes del 25 de diciembre.

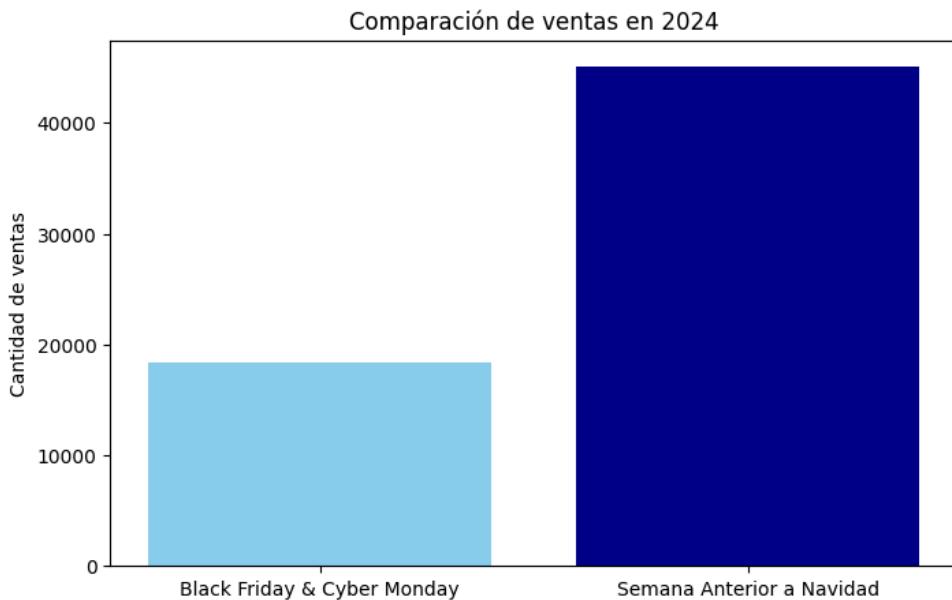


Imagen 15

## Consulta 7

En esta consulta se quiere ver la relación que hay entre el precio unitario de un producto y su descuento. ([enlace](#))

Hipótesis:

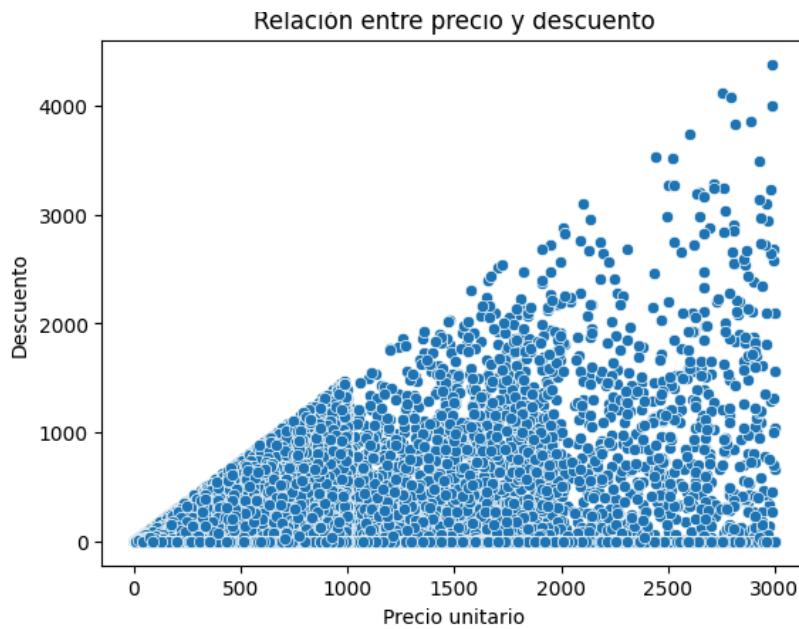
- 1) Se descartan las filas con '*unit\_price*' nulo, ya que las necesitamos para nuestro análisis y si no tienen este dato no nos aportan valor al resultado
- 2) Se descartan las filas con '*discount\_amount*' nulo, ya que las necesitamos para nuestro análisis y si no tienen este dato no nos aportan valor al resultado

Conclusión:

En el scatterplot podemos ver que se forma un triángulo y del mismo podemos sacar algunas conclusiones.

Por un lado vemos que la mayoría de productos que tienen un descuento lo tienen entre 0 y 1000.

Por otro lado, vemos que muchos productos no tienen descuento por lo que la línea de 0 está altamente poblada.



Algo para notar es que hay algunos outliers que tienen un mayor descuento que el valor del producto por lo que nos da una idea de que la columna de '*discount\_amount*' es el descuento que se le hace al producto, es decir, para obtener el precio real del producto se debe hacer:

$$\text{precio total} = \text{discount\_amount} + \text{unit\_price}$$

En un principio había pensado, erróneamente, que el precio total era el '*unit\_price*' y que el '*discount\_amount*' era el descuento sobre ese total. Esto no tiene sentido ya que si no habría productos que tienen más descuento que su valor.

## Consulta 8

En esta consulta se quiere ver la cantidad de productos robados en el primer cuatrimestre de 2024 VS primer cuatrimestre de 2025 para ver si la situación mejoró, empeoró o sigue igual. ([enlace](#))

Hipótesis:

- 1) Se dejaron solamente los robos que dan una perdida de stock, ya que los positivos se interpretaron como que se pudo recuperar la mercancía robada.
- 2) Solo consideró '*movement\_type*' del tipo '*'out'*'. Para los otros tipos considere:
  - '*'in'*' → Se recuperó la mercancía
  - '*'adjustment'*' → Hubo un error en la primera carga y se hizo un ajuste para que los datos queden correctamente
  - '*'undefined'*' → Al no saberse qué pasó no lo considero. Se prefirió tener datos confirmados.

- 3) Los valores nulos de la columna ‘order\_date’ no serán considerados ya que no se sabe en qué fecha fue la orden por lo que no puede agregarse.
- 4) Los valores de la columna ‘order\_date’ que no tengan ningún formato de DateTime serán eliminados.

Conclusión:

Como puede verse en el bar plot los robos para el primer cuatrimestre de 2024 y 2025 fueron similares, no se nota una tendencia marcada para ninguno de los dos años.

Se podría pensar que los robos en lo restante del año 2025 también serán similares a los del año 2024.

Esta información puede ser de gran utilidad para tomar medidas al respecto como contar con que se va a perder esa cantidad de stock o tomar mayores medidas de seguridad para intentar bajar la cantidad de mercancía robada.

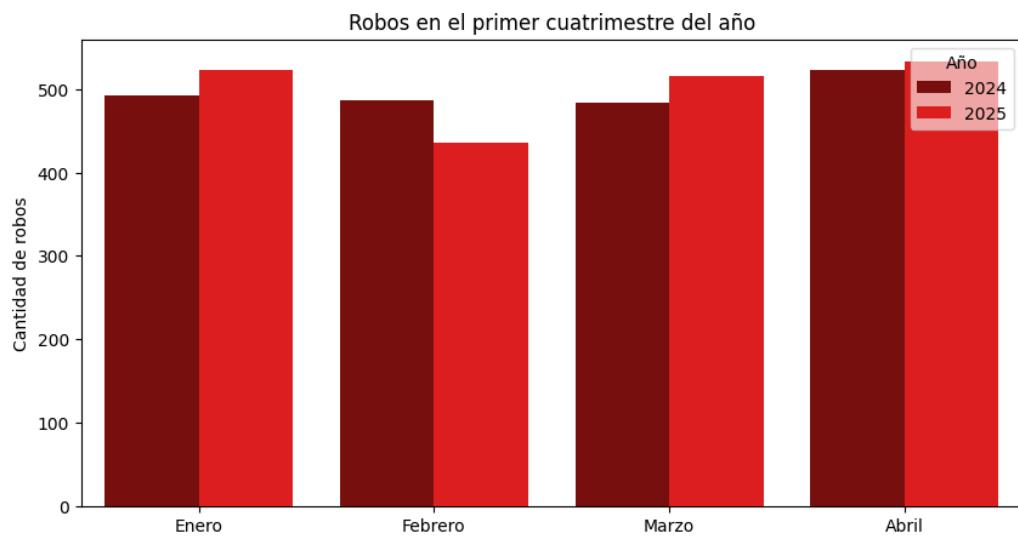


Imagen 16

## Consulta 9

Se hizo un top 10 de las categorías padre con mayor promedio de calificación por los usuarios.

El espíritu de esta consulta es poder ver qué categorías tienen mejores calificaciones para poder usar esta información en un futuro, por ejemplo para campañas publicitarias. ([enlace](#))

Hipótesis:

- 1) Descartamos las ventas que no estén calificadas

- 2) Las calificaciones son del 1 al 5. Esta condición se aplica para poder ver mejor los datos. En la actualidad las calificaciones tienen solo enteros. Si esto cambia en un futuro deberá redondearse.

Conclusión:

Podemos ver que la opiniones de las personas son mayormente favorables (4 y 5 de rating) para todas las categorías.

*Grocery & Gourmet Food, Travel y Collectibles* son las destacadas, teniendo una gran cantidad de votos de 5 puntos.

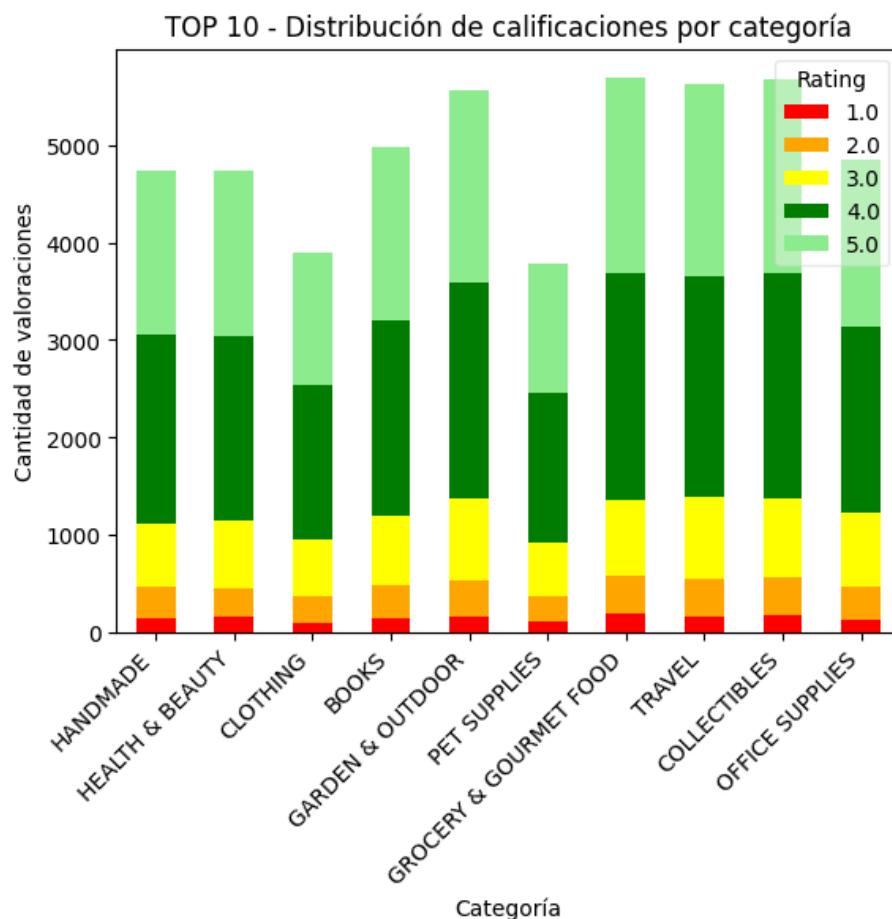


Imagen 17

Por otro lado, podemos destacar que la categoría con mayor cantidad de 1 de rating es *Garden & Outdoor*, esto puede deberse a que tiene una gran cantidad de ventas con respecto a lo demás.

Sin embargo al ver la siguiente tabla (imagen 18) vemos que esto no es así.

Por otra parte, podemos notar que el mejor promedio lo tiene Handmade aunque tiene 1000 ventas menos que el top 1.

parent_category	Cant calificaciones
GROCERY & GOURMET FOOD	5704
COLLECTIBLES	5686
TRAVEL	5629
GARDEN & OUTDOOR	5575
BOOKS	4987
OFFICE SUPPLIES	4856
HEALTH & BEAUTY	4744
HANDMADE	4740
CLOTHING	3896
PET SUPPLIES	3781

Imagen 18

## Visualizaciones

Las visualizaciones se fueron mostrando durante el desarrollo del informe aunque se dejó este apartado para que puedan verse más fácilmente.

- 1) Una continua con una línea de tiempo. ([enlace](#))
- 2) Una discreta con una continua. ([enlace 1](#) - [enlace 2](#))
- 3) Una discreta con una discreta. ([enlace 1](#) - [enlace 2](#) - [enlace 3](#))
- 4) Una continua con otra continua. ([enlace](#))
- 5) Un heatmap. ([enlace](#))
- 6) Dos visualizaciones a elección. ([enlace 1](#) - [enlace 2](#))
- 7) Una visualización hecha a mano con gráficas personalizadas:



Imagen 19

La información fue extraída de [aquí](#)

## Conclusiones

A lo largo de este trabajo se profundizó en la comprensión y el análisis del dataset provisto. Durante el proceso se identificaron diversas particularidades, entre ellas, la ausencia de una gran cantidad de *outliers*, aunque sí se detectó un gran número de registros incompletos, con valores *NaN* o *undefined*, lo que refleja problemas en la carga o disponibilidad de ciertos datos.

Asimismo, se encontraron inconsistencias en algunas columnas, como el caso de productos marcados como '*theft*' que en *inventory\_log* aparecían con valores positivos de '*stock\_quantity*', lo que generó confusión al momento de interpretarlos.

Otro aspecto relevante fue la necesidad de aplicar un proceso de estandarización previo al análisis. Varias columnas presentaban diferencias en el uso de mayúsculas, minúsculas y espacios en blanco dentro de los *strings*, lo que dificultaba el procesamiento directo de la información. Gracias a esta normalización, se logró garantizar una mayor coherencia en los datos y, por lo tanto, un análisis más confiable.

En conjunto, este trabajo no solo permitió responder las consultas planteadas, sino también adquirir experiencia en la detección, limpieza y preparación de datos, etapas fundamentales en cualquier proyecto de análisis exploratorio.